Speech for Social Good Workshop

Highly Intelligible Speech Synthesis for Spinal Muscular Atrophy Patients Based on Model Adaptation

Takuma Yoshimoto[1], Ryoichi Takashima[1], Chiho Sasaki[2], Tetsuya Takiguchi[1]

[1] Kobe University

[2] Kumamoto Health Science University

What is Spinal Muscular Atrophy ?

Spinal Muscular Atrophy (SMA)

— a type of lower motor neuron disease caused by lesions of motor nerve cells in the spinal cord

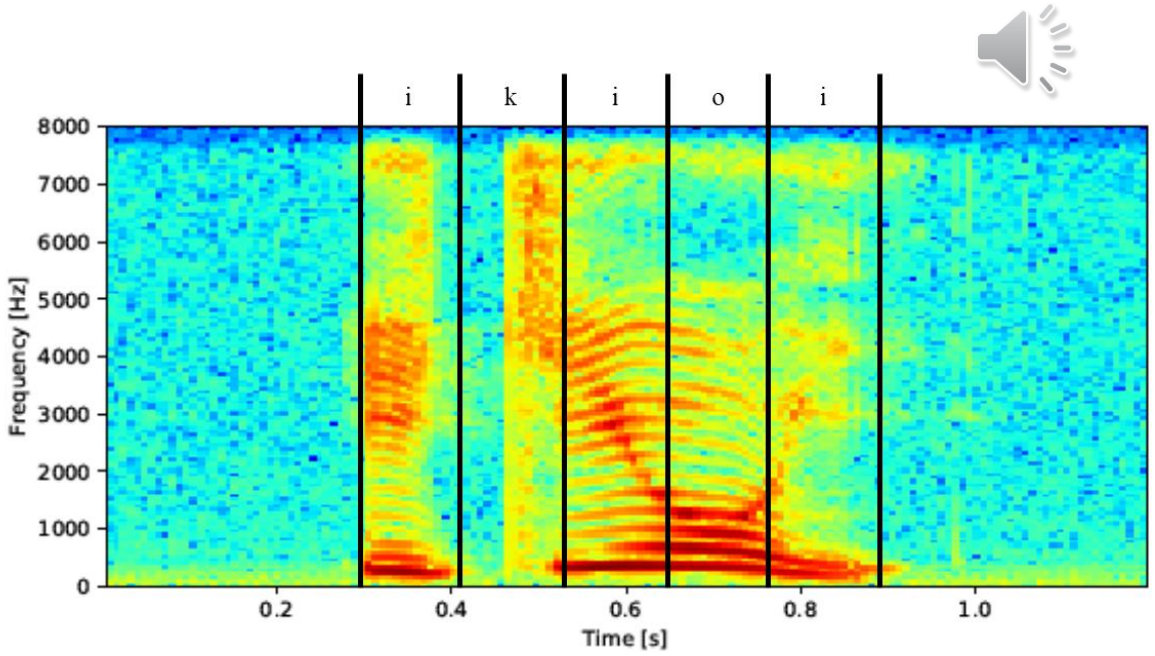
SMA type (alternate name)	Age of Onset
Type Ⅰ (Werdnig-Hoffmann disease)	Before 6 months
Type Ⅱ (Dubowitz disease)	6 – 18 months
Type Ⅲ (Kugelberg-Welander disease)	After 18 months (childhood)
Type Ⅳ	adulthood

Particularly, SMA type Ⅰ can cause **dysarthria**, dysphagia, and breathing problems

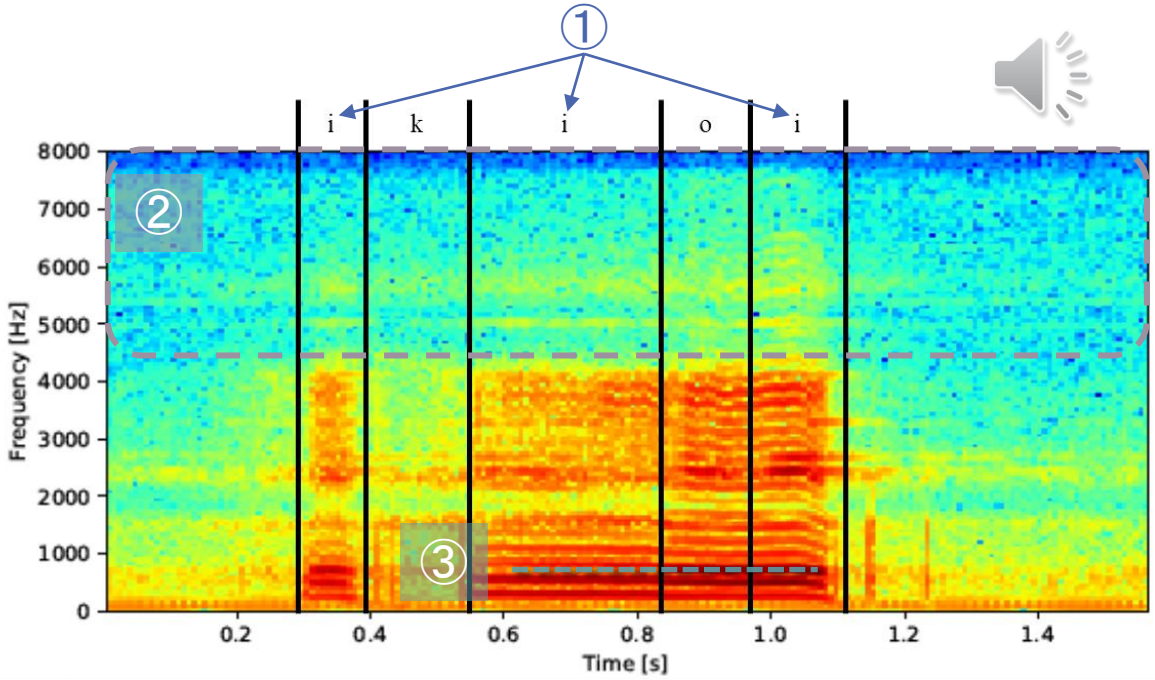
In this study, we target the speech of SMA type Ⅰ

Differences in voices (1/3)

- Spectrograms for “i k i o i (勢い)” in a healthy subject (left) and a person with SMA (right)
  - ① The duration of each phoneme is not constant
  - ② The power of the high-frequency component is weak
  - ③ Changes in the vowel are not clear



a healthy subject



a person with SMA

Differences in voices (2/3)

- Automatic Speech Recognition (ASR) results

  We examined isolated word recognition by GMM-HMM model.
  Models were trained for each phoneme, and when recognizing a word,
  the system selected the most appropriate word from a prepared dictionary.
  (vocabulary size = 216 words)

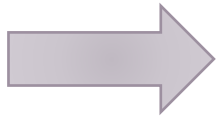
	Healthy subject	SMA patient
Accuracy	100%	15.41%

The voice of a healthy person could be recognized perfectly,
but the voice of SMA could not be recognized very correctly

Differences in voices (3/3)

- From the comparison of spectrograms, the speech of a person with SMA has the following characteristics:
  - ① The duration of each phoneme is not constant
  - ② The power of the high-frequency component is weak
  - ③ Changes in the vowel not being clear
- Automatic Speech Recognition (ASR) results indicate that ASR model cannot recognize the speech of an SMA patient very well unlike that of a healthy subject.

It is difficult to understand the speech of a person with SMA
||
barrier to communication



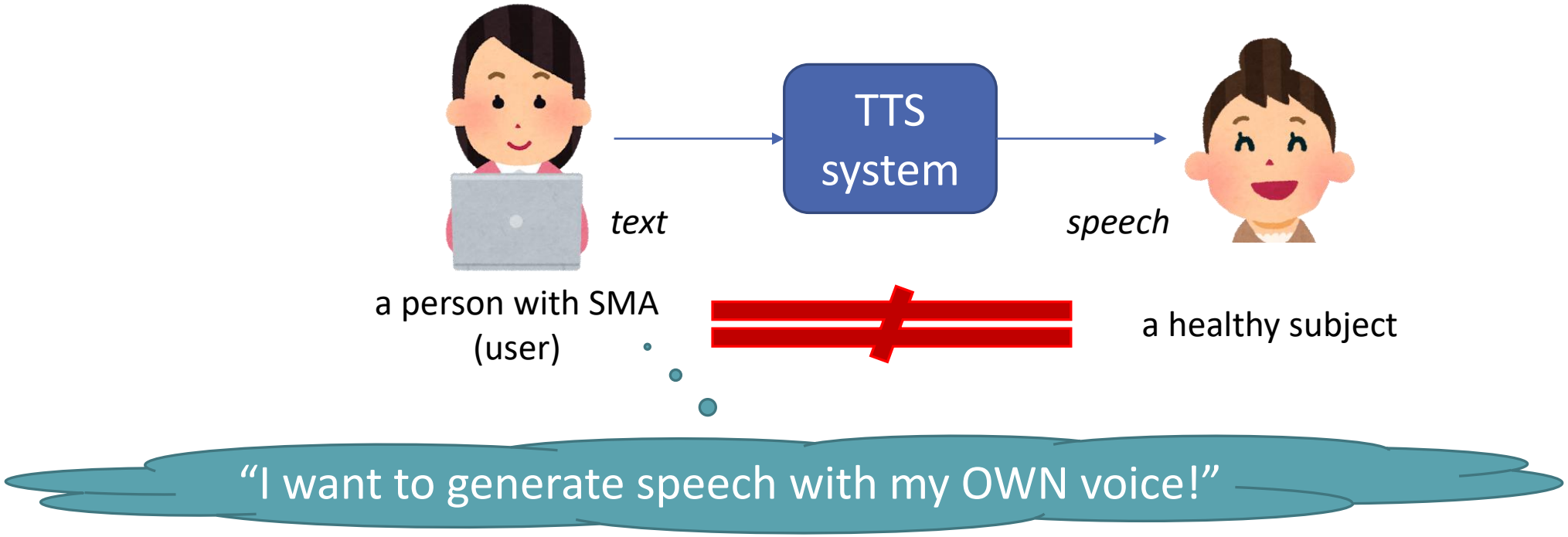
text-to-speech (TTS) system to aid in communication

Challenges of TTS in communication support

current TTS applications to aid in communication

→ train model with a healthy subject, not with the data of the user

→ synthesized speech is completely different from the user's own



Creating TTS for a person with SMA (1/2)

A possible way:

"record the voice of the SMA patient and train a TTS model with only that data"

Problem ①

Large volume of recordings is too much of a burden on the SMA person

Purpose ①

train a TTS model with a small amount of SMA patient speech data

Problem ②

The speech would be synthesized indistinctly, just like the SMA person's original voice

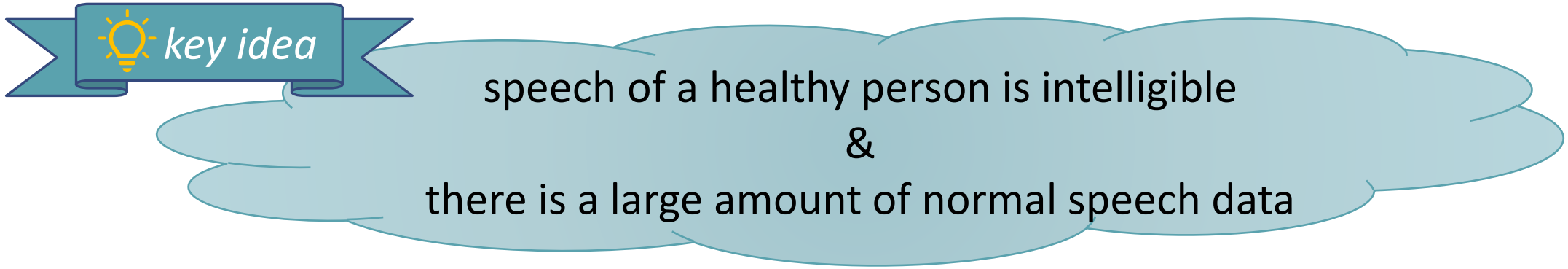
Purpose ②

synthesize speech that is both individual and intelligible

Creating TTS for a person with SMA (2/2)

train a TTS model with a small amount of SMA patient speech data

synthesize speech that is both individual and intelligible



PROPOSED METHOD

After creating a TTS model using a large amount of a healthy subject's speech, the model is adapted using a small amount of an SMA patient's own speech

Proposed Method

- **Pre-training**…make a model for a healthy subject

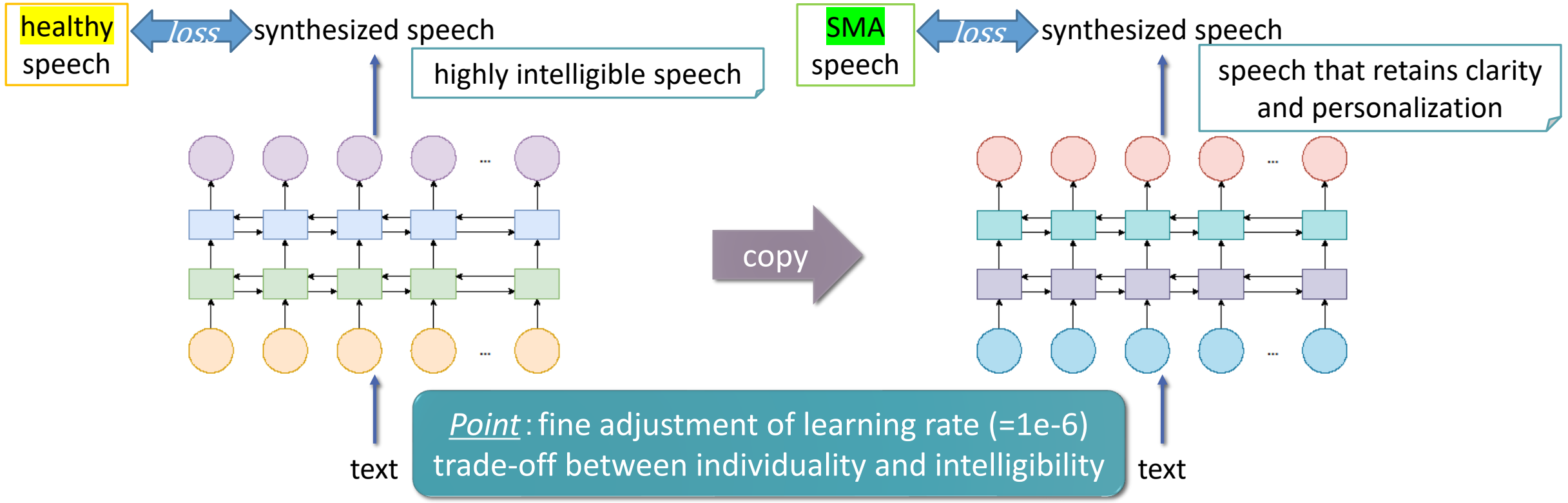
Step 1. Random initial values for the TTS model

Step 2. Training with a large amount of healthy data

- **Fine-tuning**…adapt the model for an SMA patient

Step 1. Copying a TTS model for a healthy subject

Step 2. Training with a small amount of SMA data



Experimental conditions (1/2)

- Speech data

  healthy person: 503 phoneme-balanced **sentences** were uttered by one Japanese female

  SMA person: 215 phoneme-balanced **words** were uttered by one Japanese female (SMA type Ⅰ)

  Each word was uttered 4-5 times repeatedly (total 1,070 utterances)

- Model structure

  we use two models: an acoustic model and a duration model

  three layers of bidirectional LSTM (long short-term memory) having 1,024 cells in each layer

  only the acoustic model is adapted

- Vocoder

  WORLD

Experimental conditions (2/2)

- Acoustic features
  - 60-dimensional melcepstrum
  - a band aperiodicity parameter (BAP)
  - a logarithmic fundamental frequency (F0)

  } + Δ + ΔΔ

  - a voiced/unvoiced flag

  total dimension: 187

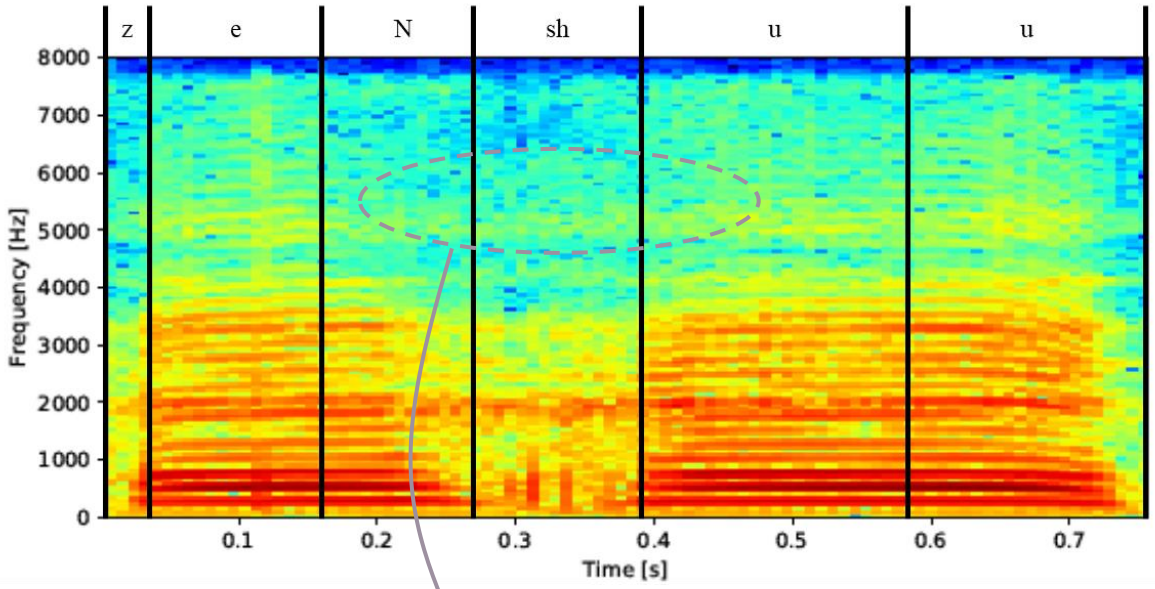
- Other conditions
  - sampling rate: 16 kHz frame shift: 5 msec
  - number of dimensions for the linguistic features: 975

- Evaluation method
  - Comparison of spectrograms
  - Subjective evaluation experiment (intelligibility and individuality)

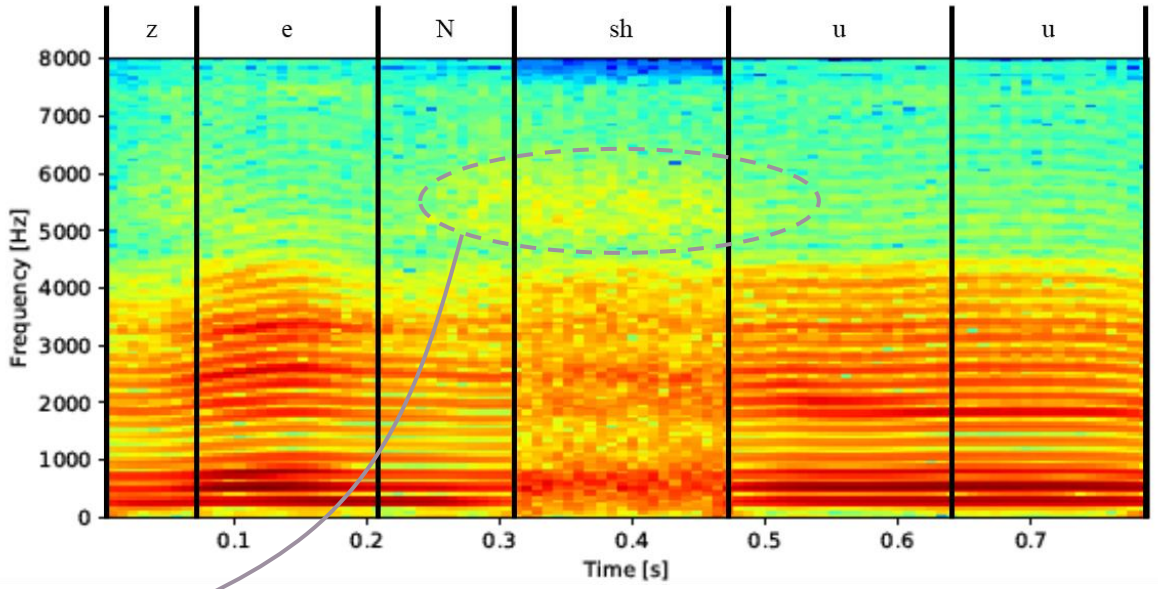
Experimental results (1)

Comparison of spectrograms “z e N sh u u (全集)”

〈original speech〉



〈synthesized speech〉



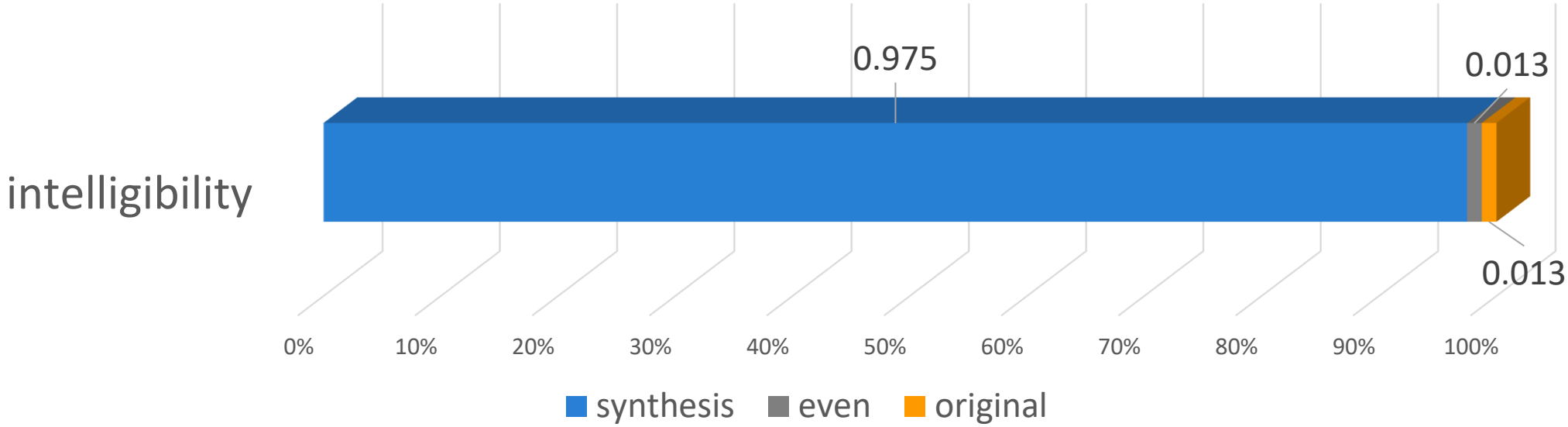
/sh/ is a fricative sound, which has strong high-frequency components

emphasized by speech synthesis

improvement of intelligibility

Experimental results (2-1)

Subjective evaluation (AB test)

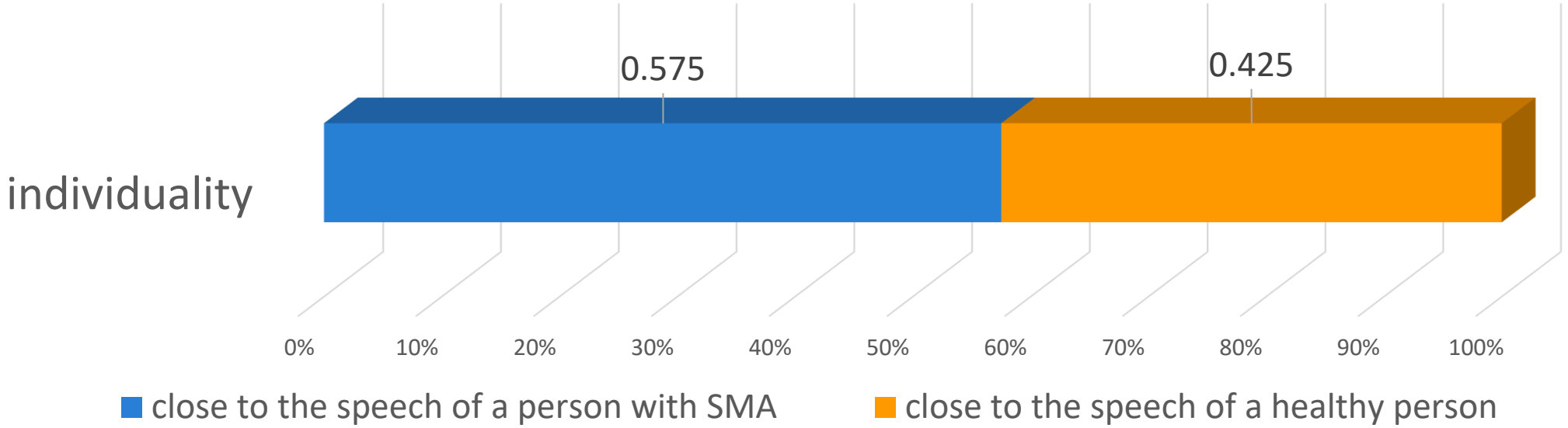


Almost all (97.5%) answered that the synthesized speech was easier to hear than the original one

- ✓ Consonants were heard more clearly

Experimental results (2-2)

Subjective evaluation (XAB test)



The synthesized speech was predominantly close to the speech of a person with SMA, at around 60%

- ✓ Proposed method achieved to synthesize many speeches which are similar to the SMA speech
- ✓ Because we applied weak adaptation to ensure intelligibility, this might have a negative effect on the individuality

e.g. “z e n sh u u (全集)” 〈SMA original speech〉 〈speech of a healthy subject〉 〈synthesized speech〉

Conclusion and future works

Conclusion

Speech synthesis for SMA based on speaker adaptation of a healthy TTS model can produce synthesized speech with improved intelligibility while maintaining individuality.

Future works

- investigate methods for solving the trade-off problem between intelligibility and individuality
- investigate objective evaluation criteria that can quantitatively evaluate the intelligibility and individuality of synthesized speech (e.g., speech/speaker recognition accuracy)

Thank you for listening!